ProDiff: Progressive Fast Diffusion Model for High-Quality Text-to-Speech

Year: 2022 | Citations: 227 | Authors: Rongjie Huang, Zhou Zhao, Huadai Liu, Jinglin Liu, Chenye Cui

---

Abstract

Denoising diffusion probabilistic models (DDPMs) have recently achieved leading performances in many generative tasks. However, the inherited iterative sampling process costs hinder their applications to text-to-speech deployment. Through the preliminary study on diffusion model parameterization, we find that previous gradient-based TTS models require hundreds or thousands of iterations to guarantee high sample quality, which poses a challenge for accelerating sampling. In this work, we propose ProDiff, on progressive fast diffusion model for high-quality text-to-speech. Unlike previous work estimating the gradient for data density, ProDiff parameterizes the denoising model by directly predicting clean data to avoid distinct quality degradation in accelerating sampling. To tackle the model convergence challenge with decreased diffusion iterations, ProDiff reduces the data variance in the target site via knowledge distillation. Specifically, the denoising model uses the generated mel-spectrogram from an N-step DDIM teacher as the training target and distills the behavior into a new model with N/2 steps. As such, it allows the TTS model to make sharp predictions and further reduces the sampling time by orders of magnitude. Our evaluation demonstrates that ProDiff needs only 2 iterations to synthesize high-fidelity mel-spectrograms, while it maintains sample quality and diversity competitive with state-of-the-art models using hundreds of steps. ProDiff enables a sampling speed of 24x faster than real-time on a single NVIDIA 2080Ti GPU, making diffusion models practically applicable to text-to-speech synthesis deployment for the first time. Our extensive ablation studies demonstrate that each design in ProDiff is effective, and we further show that ProDiff can be easily extended to the multi-speaker setting.